

Exploring the Energy Landscapes of Protein Folding Simulations with Bayesian Computation

Nikolas S. Burkoff,[†] Csilla Várnai,[†] Stephen A. Wells,[‡] and David L. Wild^{†*}

[†]Systems Biology Centre and [‡]Department of Physics and Centre for Scientific Computing, University of Warwick, Coventry, United Kingdom

ABSTRACT Nested sampling is a Bayesian sampling technique developed to explore probability distributions localized in an exponentially small area of the parameter space. The algorithm provides both posterior samples and an estimate of the evidence (marginal likelihood) of the model. The nested sampling algorithm also provides an efficient way to calculate free energies and the expectation value of thermodynamic observables at any temperature, through a simple post processing of the output. Previous applications of the algorithm have yielded large efficiency gains over other sampling techniques, including parallel tempering. In this article, we describe a parallel implementation of the nested sampling algorithm and its application to the problem of protein folding in a Gō-like force field of empirical potentials that were designed to stabilize secondary structure elements in room-temperature simulations. We demonstrate the method by conducting folding simulations on a number of small proteins that are commonly used for testing protein-folding procedures. A topological analysis of the posterior samples is performed to produce energy landscape charts, which give a high-level description of the potential energy surface for the protein folding simulations. These charts provide qualitative insights into both the folding process and the nature of the model and force field used.

INTRODUCTION

Approximately 50 years ago, Anfinsen and colleagues (1) demonstrated that protein molecules can fold into their three-dimensional native state reversibly, leading to the view that these structures represented the global minimum of a rugged funnel like energy landscape (1–3).

According to the hierarchical folding theory of Baldwin and Rose (4,5), a protein folds by first forming local structural elements, namely, α -helices and β -strands. These secondary structure elements then interact with each other, resulting in the formation of the folded protein. The formation of local structural elements reduces the entropy of the protein (for example, the side chains of helical residues are strongly constrained by the rest of the helix). This loss of entropy is compensated by favorable short-range interactions, including hydrogen bonding and desolvation of backbone polar groups. This is considered to be a fundamental property of proteins, and any model system attempting to simulate protein folding should mimic this property.

Although there has been recent evidence of hierarchical folding in long timescale molecular dynamics simulations made possible by the use of custom designed supercomputers (6), simplified Gō-type models remain an important class of protein models in the investigation of energy landscapes. Gō models assume that nonnative interactions do not

contribute to the overall shape of the folding energy surface (7,8). In this work we use an extended Gō-type model, in which a Gō potential captures interactions between contacts of the native state of the protein, but attractive nonnative interactions are also permitted (for example, hydrogen bonds can form between residues that are not in contact in the native state). This addition allows us to explore a more realistic rugged energy landscape compared to the “perfect funnel” found in a standard Gō model (8), while maintaining the ability to perform simulations with limited computational resources.

The energy landscapes of protein-folding simulations are most commonly visualized in terms of two- or three-dimensional plots of microscopic or free energy versus a reaction coordinate, such as the fraction of residue contacts in common with the native state or the root mean-square deviation (RMSD) between a given conformation and the native state (9,10). Originally developed for reduced lattice models, these approaches have since been used for all-atom off-lattice simulations, although, in these more realistic models, they offer only an indirect visualization of the energy landscape at a single scale (11). Projection into the space defined by principal components analysis of the contact map has also been used to provide a two-dimensional visualization of the energy surface (12). Techniques adapted from robotic motion planning have been used to provide a probabilistic roadmap of protein folding, which may be mapped onto a conceptual drawing of the potential energy surface (13). Protein potential energy surfaces and folding funnels have also been visualized by disconnectivity graphs (14) and scaled disconnectivity graphs (15,16). Although these latter methods have the advantage of providing a visualization of the whole energy landscape, they rely on creating a large database of local energy

Submitted September 13, 2011, and accepted for publication December 27, 2011.

*Correspondence: d.l.wild@warwick.ac.uk

This is an Open Access article distributed under the terms of the Creative Commons-Attribution Noncommercial License (<http://creativecommons.org/licenses/by-nc/2.0/>), which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editor: Michael Levitt.

© 2012 by the Biophysical Society. Open access under CC BY-NC-ND license.
0006-3495/12/02/0878/9

doi: 10.1016/j.bpj.2011.12.053

minima of the surface, and are therefore impractical for large systems; they also do not provide information about the entropy of the system (which governs the widths of the conceptual protein-folding funnel).

The funnel like nature of the energy landscape provides a challenging conformational space for computer simulations to explore, because only an exponentially small number of conformations have low energy and low entropy and are found toward the bottom of the funnel; the system also undergoes a first-order phase transition as the protein collapses into its native state. In this work, we use nested sampling to explore the energy landscapes of protein folding simulations. Nested sampling is a Bayesian sampling technique introduced by Skilling (17,18), designed to explore probability distributions where the posterior mass is localized in an exponentially small area of the parameter space. It both provides an estimate of the evidence (also known as the marginal likelihood, or partition function) and produces samples of the posterior distribution. Nested sampling offers distinct advantages over methods such as simulated annealing (19), Wang-Landau sampling (20), parallel tempering (replica exchange) (21), and annealed importance sampling (22), in systems characterized by first-order phase transitions (17,23). The technique reduces multidimensional problems to one dimension and has a single key parameter in the trade-off between cost and accuracy. The calculation of free energies by thermodynamic integration (24) and thermodynamic observables, such as heat capacities, typically involves multiple simulations at different temperatures. Nested sampling provides an efficient framework for computing the partition function and hence thermodynamic observables at any temperature, without the need to generate new samples at each temperature. Hence, it allows us to directly investigate the macroscopic states of the protein-folding pathway and evaluate the associated free energies. Nested sampling has previously been used in the field of astrophysics (25) and for exploring potential energy hyper-surfaces of Lennard-Jones atomic clusters (23), yielding large efficiency gains over parallel tempering. Its use in this article represents, to our knowledge, the first application of this technique to a biophysical problem.

MATERIALS AND METHODS

In general, the energy of a polypeptide, $E(\Omega, \theta)$, is defined by its conformation, Ω , and arbitrary interaction parameters, θ . These interaction parameters may be as diverse as force constants, distance cutoffs, dielectric permittivity, atomic partial charges, etc. This energy, in turn, defines the probability of a particular conformation, Ω , at inverse thermodynamic temperature β via the Boltzmann distribution

$$P(\Omega, \theta | \beta) = \frac{1}{Z(\theta, \beta)} \exp[-E(\Omega, \theta)\beta], \quad (1)$$

$$Z(\theta, \beta) = \int d\Omega \exp[-E(\Omega, \theta)\beta], \quad (2)$$

where $Z(\theta, \beta)$ is the partition function (or evidence, in Bayesian terminology). In the following, energy is expressed in units of RT , the product of the molar gas constant and absolute temperature and $\beta = 1/RT$.

In Bayesian statistics, with θ an unknown parameter, D the observed data, and H the underlying model or hypothesis, we have the following relation (Bayes' rule)—posterior \times evidence = likelihood \times prior—

$$\mathcal{P}(\theta|D, H)Z = \mathcal{P}(D|H, \theta)\mathcal{P}(\theta|H),$$

where Z , the evidence, is defined as

$$Z = \int \mathcal{P}(D|H, \theta)\mathcal{P}(\theta|H)d\theta.$$

Nested sampling provides an algorithm for estimating the evidence, $Z = \mathcal{P}(D|H)$, and the procedure additionally explores the posterior distribution, allowing its properties to be estimated.

Procedure

We define $X(\lambda) = \lambda$ to be the proportion of the prior distribution with likelihood $L(X) > \lambda$. Then, following Skilling (17), the evidence is

$$Z = \int_0^1 L(X)dX,$$

where $L(X(\lambda)) = \lambda$ and $dX = \pi(\theta)d\theta$, with $\pi(\theta)$ the prior distribution. Fig. 1 shows the graph of L against X (this is not to scale, as normally the bulk of the posterior is in an exponentially small area of the phase space). L is a decreasing function of X , as the restriction on the likelihood becomes tighter as λ increases. The area under the curve is Z . The nested sampling procedure estimates points on this curve (see Algorithm, below) and then uses numerical integration to calculate Z .

Algorithm

1. Sample (uniformly, with respect to the prior distribution) K points of the parameter space $\{\theta_1 \dots \theta_K\}$, i.e., the “active list”; then calculate their likelihoods: $\{L(\theta_1), \dots, L(\theta_K)\}$.
2. Take the sample point with the smallest likelihood; save it as (L_1, X_1) (see below for an estimate of X); remove this point from the active list.

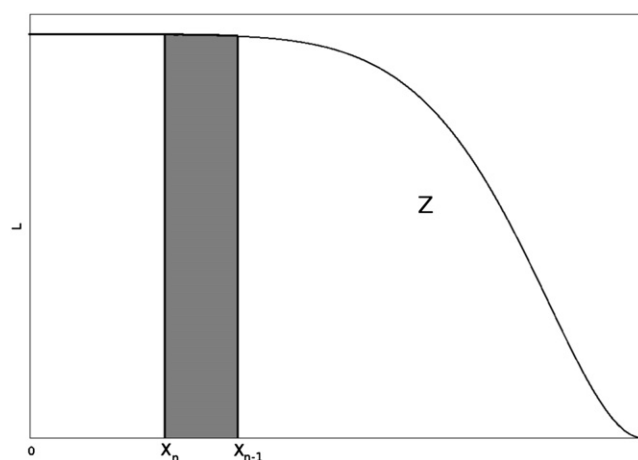


FIGURE 1 Evidence Z is the area under the function $L(X)$. The sample θ_n represents $X_{n-1} - X_n$ of the phase space volume; the proportion of the x axis is shaded. Its weighting for the posterior is $L_n (X_{n-1} - X_n)/Z$; the proportion of Z is shaded.

3. Generate a new point θ sampled uniformly (with respect to the prior distribution) from those points with likelihood $L(\theta) > L^* = L_1$; then add it to the active list.
4. Repeat Steps 2 and 3, generating $(L_2, X_2), (L_3, X_3), \dots, (L_i, X_i), \dots$

X_i is located at the largest of N numbers uniformly distributed on $(0, X_0)$, where $X_0 = 1$. Skilling (17) suggests using the expected value of the shrinkage ratio, X_i/X_{i-1} , to estimate X_i (the estimate of X for iteration i), where X_i is the largest of N numbers uniformly distributed on $(0, X_{i-1})$. The shrinkage ratio has the probability density function $f(t) = Kt^{K-1}$, with mean and standard deviation $\log(t) = (-1 \pm 1)/K$, and, as each shrinkage ratio is independent, we find, if uncertainties are ignored,

$$\log(X_i) = (-i \pm \sqrt{i})/K \Rightarrow X_i \approx \exp(-i/K).$$

It is also possible to use the arithmetic expected value to estimate X_i (26). This implies that $X_i = \alpha^i$, where $\alpha = K/(K+1)$. In the limit of large K , these two approaches are identical and henceforth we will use $\alpha = \exp(-1/K)$ or $K/(K+1)$, and $X_n = \alpha^n$.

Parallel nested sampling

For high-dimensional systems, sampling uniformly (conditional upon the likelihood being above a fixed value, L^*) is not computationally tractable. In this case, a Markov chain can be used to explore the parameter space (22). To generate a new point, one of the active set of points (not necessarily the one with the lowest likelihood) is chosen to be the start of a short Monte Carlo (MC) run, with all moves that keep the likelihood above L^* being accepted.

Starting the MC run from a copy of one of the points of the active set, chosen at random, is crucial to nested sampling. Suppose we have a bimodal likelihood function. Once L^* is sufficiently high, the region of the parameter space the chain is allowed to explore will no longer be connected; it will have two disconnected components. Without copying, all active points that enter the subordinate component will be trapped there. With copying, provided at least one enters the dominant mode, then as L^* increases, active points in the subordinate mode will be replaced by ones from the dominant mode. This is particularly important for likelihood functions for which the dominant mode splits again at a higher likelihood. In general, for a given K , if the relative phase space volume of a mode is $< 1/K$ in comparison to the rest of the space at the splitting likelihood, the chances of nested sampling exploring the mode is small (23). Therefore, the parameter K controls the resolution of the exploration.

The number of trial MC moves per nested sampling iteration, m , is another key parameter when using nested sampling for higher dimensional systems. If m is too small, the parameter space is inadequately explored; new active set samples and the current conformations they are copied from remain very similar. Setting m too high results in longer than necessary runtimes, as conformations partway through the MC run are already sufficiently different from their starting positions. Hence, K controls which regions of the parameter space are available to the algorithm and m controls how well these regions are explored.

We parallelized the nested sampling algorithm by removing the P points with the lowest likelihood at each nested sampling iteration, one for each processor used. Each processor then runs its own independent MC simulation to replace one of the removed points. For post processing, at each iteration we only store the point that has the P^{th} lowest likelihood and adjust α accordingly; $\alpha = 1 - P/(K+1)$.

Running a parallel nested sampling algorithm with K points explores the parameter space more effectively than P serial nested sampling simulations each with K/P points in the active set, while requiring equal computational resources. Consider a likelihood function, which splits n times in the dominant mode (i.e., contains the majority of the evidence), with the probabilities of an exploratory active point falling into the dominant mode being

W_1, W_2, \dots, W_n at the critical likelihood (which is the likelihood of splitting). Defining success as exploring the dominant mode at the n^{th} split in at least one simulation, it can be shown, using an argument similar to that of Sivia and Skilling (18), that

$$\mathbb{P}(\text{success} | \text{one simulation with } K \text{ points}) = \Pi_g \left[1 - (1 - W_g)^K \right] \quad (3)$$

and

$$\begin{aligned} \mathbb{P}(\text{success} | P \text{ simulations with } K/P \text{ points}) \\ = 1 - \left(1 - \Pi_g \left(1 - (1 - W_g)^{K/P} \right) \right)^P. \end{aligned}$$

For example, if $n = 2$, $W_1 = W_2 = 0.1$, $K = 32$, and $P = 4$, then $\mathbb{P}(\text{success} | \text{parallel}) = 0.933$ and $\mathbb{P}(\text{success} | \text{serial}) = 0.792$.

Posterior samples

The sample points removed from the active set, labeled $\theta_1, \theta_2, \dots$, say, can be used to estimate properties of the posterior distribution. Sample point θ_n represents

$$\omega_n = X_{n-1} - X_n$$

of the phase space volume (with respect to the prior distribution) and hence

$$\chi_n = \frac{(X_{n-1} - X_n)L(\theta_n)}{Z}$$

is the relative volume of the posterior space that θ_n represents; see Fig. 1.

In the case of a Boltzmann distribution, at inverse temperature β , $L(\theta_n) = \exp(-E_n\beta)$ and hence, by calculating $\chi_n(\beta)$, a single nested sampling simulation can provide the expectation value of any thermodynamic observable, such as heat capacity, at any temperature. Given a property $Q(\theta|\beta)$ of the posterior,

$$\mathbb{E}(Q|\beta) \approx \sum_i \chi_i(\beta) Q(\theta_i). \quad (4)$$

In energetic terms, the nested sampling scheme is built from a set of decreasing energy levels, $\{E_n\}$, with the energy of conformation Ω_n given by Eq. 5. Each energy level has an associated weight, which is also decreasing. At each energy level, a set of K sample points (or conformations), $\{\Omega_n^i\}$, is obtained by uniform sampling from the energy landscape below E_n : $\Omega_n^i \sim U(\Omega: E(\Omega) < E_n)$. After every iteration, a new lowest energy level E_{n+1} is defined to be at a fixed fraction, α , of the current energy distribution. In this way, a fraction α^n of the whole phase space has energy below E_n and a fraction α^{n+1} has energy below E_{n+1} . The phase space volume will therefore shrink exponentially, by a factor of α , with every nested sampling iteration, and the algorithm is able to locate exponentially small regions of phase space.

The protein model

The polypeptide model we use is adapted from our previous published work (27–30). It is fully described in the Supporting Material and a summary is provided below.

Our polypeptide model features all-atom representations of the polypeptide backbone and β -carbon atoms. Other side-chain atoms are represented by one or, in the case of branched side chains, two pseudoatoms, following Srinivasan and Rose (31).

For a given protein sequence, R , the Boltzmann distribution defines the probability, $P(R, \Omega|\beta)$, of it adopting a particular conformation, Ω , at

inverse thermodynamic temperature β . This probability can be factorized into the product of the sequence-dependent likelihood for a given conformation and the prior distribution of conformations, $P(R, \Omega) = P(R|\Omega)P(\Omega)$. This can be rewritten in energetic terms as

$$E(R, \Omega) = -\ln P(R|\Omega) + E(\Omega), \quad (5)$$

where sequence-dependent and sequence-independent contributions to the energy are separated. We assume that the sequence-independent term, $E(\Omega)$, is defined by short-range interactions among the polypeptide backbone, β -carbon, and pseudo-atoms. At room temperature, van der Waals repulsions and covalent bonding between atoms are extremely rigid interactions that contribute to this energy. Another large contribution comes from hydrogen bonding, but the magnitude of this interaction is vaguely understood. The sequence-dependent part of the potential (the negative log-likelihood) can be approximated by the pairwise interactions between side chains, which make the largest contribution to this term. In this work, these interactions are modeled by a Gō-type potential based on a regularized native contact map (27), which contains lateral contacts in parallel and anti-parallel β -sheets and contacts between residues i and $i + 3$ in α -helices (32,33). Our model also includes a hydrophobic packing term; hydrophobic side chains coming into contact with hydrophobic or amphipathic side chains are rewarded with a decrease in energy (31). The force constants for these side-chain interactions, as well as backbone hydrogen bonding, are optimized using a novel statistical machine learning technique (29).

Nested sampling is initialized with K conformations, uniformly distributed over the space of dihedral angles (i.e., every $\phi_i, \psi_i \sim U[-180^\circ, 180^\circ]$). To generate new sample points we use our implementation of an efficient Metropolis Monte Carlo (MMC) algorithm (28,30), which relies on local Metropolis moves, as suggested in earlier studies (34). In contrast to other programs that rely on local Metropolis moves in the space of dihedral angles, our sampler utilizes local crankshaft rotations of rigid peptide bonds in Cartesian space. An important feature of our model is the elasticity of the α -carbon valence geometry. With flexible α -carbon valence angles, it becomes possible to use crankshaft moves inspired by earlier MMC studies of large-scale DNA properties. The amplitudes of proposed crankshaft rotations were chosen uniformly from $[-\alpha_0, \alpha_0]$ where, at every 2000 nested sampling iterations, α_0 (the maximum allowed proposed amplitude) was recalculated, attempting to keep the acceptance rate at 50% (the trial MC moves used for this calculation were then ignored).

We ran simulations until $Z(\beta)$ converges for $\beta = 1$ ($T = 25^\circ\text{C}$), which implies that we have sampled from the thermodynamically accessible states for all temperatures smaller than β ($>T$). The nested sampling algorithm marches left across the x axis of Fig. 1. The step size is constant in $\log X$ and the larger the K , the smaller the step size. For a given protein and β , we find that simulations terminate at approximately the same point on the x axis (for protein G, with $\beta = 1$, this is $\sim e^{-440}$). This implies that the total number of iterations is proportional to K , and the total number of MC moves is proportional to mK . The results for protein G shown below are from a simulation with $K = 20,000$ and $m = 15,000$, which used 32 processors (Xeon X5650; Intel, Santa Clara, CA), had 1.38×10^{11} MC moves, and took ~ 22 h.

Energy landscape charts

We use the algorithm recently introduced by Pártay et al. (23), which uses the output of a nested sampling simulation to generate an energy landscape chart, facilitating a qualitative understanding of potential energy surfaces. It has the advantage of showing the large-scale features of the potential energy surface without requiring a large number of samples.

The output of a nested sampling simulation is a sequence of sample points with decreasing energy. Each sample point (conformation), Ω_n , represents $\omega_n = \alpha^{n-1} - \alpha^n$ of the phase space and has energy $E_n(\Omega_n)$. A metric defining the distance between two conformations is required, and using this, a topological analysis of the sample points is performed. As the

metric, we use the root mean-square deviation of the backbone and side-chain nonhydrogen atoms of a pair of conformations; that is, the sum of the Euclidean distances of corresponding atoms after the two conformations have been translated and rotated in space to minimize the overall distance.

A graph \mathcal{G} is created with the sample points as nodes and arcs joining a sample to the k nearest samples that have higher energy. In this work, k is chosen to be 15 throughout. We then start with an empty graph (\mathcal{G}'), adding nodes one at a time (starting with the lowest energy) to gradually rebuild \mathcal{G} .

Energy landscape charts are produced with energy on the vertical axis, and, at a given energy E_n , the width of the chart is proportional to the sum of the weights of the points below that energy (i.e., $\omega_n + \omega_{n+1} + \dots$), that is, the available phase space volume in the prior space, contained at $<E_n$. On the horizontal scale, the chart is split into different basins corresponding to the disconnected subgraphs that exist when sample n is added to \mathcal{G}' . The relative widths of the basins is given by the ratio of the sum of the weights of the sample points in the disconnected subgraphs. The ordering of the basins horizontally is arbitrary. Due to the rapid shrinking of the available phase space volume with decreasing energy, for better visualization, a horizontal scaling is applied by an exponential function of the energy, similar to Pártay et al. (23). The energy landscape chart represents a potential energy landscape for the system.

We also use a variant of the energy landscape charts where the width of the chart is proportional to the sum of the posterior weights, $\chi_n = \omega_n \exp(-E_n\beta)/Z(\beta)$, i.e., $(\chi_n + \chi_{n+1} + \dots)$, at inverse temperature β . Hence, the relative widths of the basins correspond to the probabilities of adopting a conformation from one basin or another. These energy landscape charts, therefore, represent the energy landscape as it is experienced by the protein at inverse temperature β . In the following, the two versions will be referred to as prior and posterior energy landscape charts, according to the weights used in the calculation of their basin widths.

RESULTS

To validate the nested sampling procedure, we simulated the folding of an isolated 16-residue polyalanine β -hairpin. We then conducted folding simulations on a number of small proteins that are commonly used for testing protein folding procedures: protein G (PDB code 1PGA), the SH3 domain of Src tyrosine kinase (PDB code 1SRL), and chymotrypsin inhibitor 2 (PDB code 2CI2).

Isolated polyalanine β -hairpin

We used a Gō-like potential to simulate the folding of an isolated 16 residue polyalanine β -hairpin. Fig. S1 in the Supporting Material (bottom panel) shows a snapshot of five (equally spaced along the $\log(X)$ axis) conformations from a single simulation with $K = 1000$, $m = 2500$ (a total of 1.12×10^8 MC moves). At the beginning there is a rapid decrease in energy, moving from extended conformations (at first those with van der Waals collisions) to hairpinlike structures (A–C). The final part of the simulation moves through the exponentially small volume of the phase space containing hairpinlike structures, gradually decreasing in energy toward a fully formed hairpin (D and E).

We used the hairpin to check the behavior of the nested sampling procedure: Fig. S1 (top panel) shows how α_0 (the maximum proposed crankshaft rotation amplitude) varies with the energy threshold for a simulation with

$K = 1000$. As lower energy is reached, α_0 is reduced to keep the acceptance rate near 0.5. Fig. S1 (second panel) shows the acceptance rate. Fig. S1 (third panel) shows the difference between the start and end points of a single MC chain, specifically the drift per dihedral angle, where the drift is the L_2 -norm of the dihedral angles.

The protein model used stabilizes room temperature secondary structure formation; it folds isolated helices and hairpins very effectively. This is reflected in the energy landscape charts that consist of a single funnel (not shown).

Fig. S2 (top) shows the time evolution of the dihedral angles of four residues of the 16-residue polyalanine. The formation of the hairpin can be clearly seen. For example, the dihedral angles of the residues in the strands 4 and 11 converge to the standard β -sheet area of the Ramachandran plot. The Gō-like potential used was designed for a hairpin with a two-residue turn, and this is found to be the case. The dihedral angles of the turn residues 8 (60 ± 15 , -90 ± 30) and 9 (-150 ± 30 , 0 ± 30) are closest to the values of a type II' turn ((60, -120) and (-80 , 0)) (35). Fig. S2 (bottom) shows the energy of the snapshots (right-hand axis) for nested sampling plotted against time. The monotonic decrease of the energy over a very large energy range allows us to view the formation of the hairpin.

Due to the nature of the model used, the folding pathway of the hairpin is relatively simple to sample, and parallel tempering can also successfully fold the hairpin. However, in this case, we need a very large temperature range to explore the whole parameter space and view the folding pathway in its entirety. For example, Fig. S2 (bottom) shows the energy of two of the parallel tempering chains; room temperature and 300°C. For real proteins, which have more complicated energy landscapes and possibly high energy barriers, it is difficult to know the temperature range

required for parallel tempering to explore the entire parameter space and not be trapped in a particular basin. Nested sampling, with its top-down, temperature-independent approach, does not suffer from this problem.

Another of the advantages of nested sampling is that simulations are temperature-independent, and hence can provide estimates of thermodynamic variables at any temperature. Fig. 2 shows the heat capacity (C_v) curve for the 16-residue polyalanine. The curves were calculated using nested sampling (converged down to -25°C , so that the C_v curve does not stop abruptly at room temperature), and parallel tempering. The solid line is calculated using 10 nested sampling simulations each with 1.3×10^9 MC moves. The dashed lines show twice the standard error. The parallel tempering curve shows the heat capacity using 10 parallel tempering simulations (again each with 1.3×10^9 MC moves) with error bars showing twice the standard error. For parallel tempering, the heat capacity is only calculated for discrete temperatures and a procedure such as Boltzmann reweighting (36) is needed to calculate the continuous curve.

There appears to be good agreement between the methods. Previous results have found nested sampling to be more efficient at calculating the heat capacity curves (23). In this example, we found nested sampling to be of similar efficiency to parallel tempering. We believe this to be because, unlike the system presented in Pártay et al. (23), our phase transition (from coil to hairpin) occurs over a very large energy (and hence temperature) range from which parallel tempering can successfully sample.

Protein G

Protein G is a 56-residue protein consisting of an antiparallel four-stranded β -sheet and an α -helix, with a β -Grasp

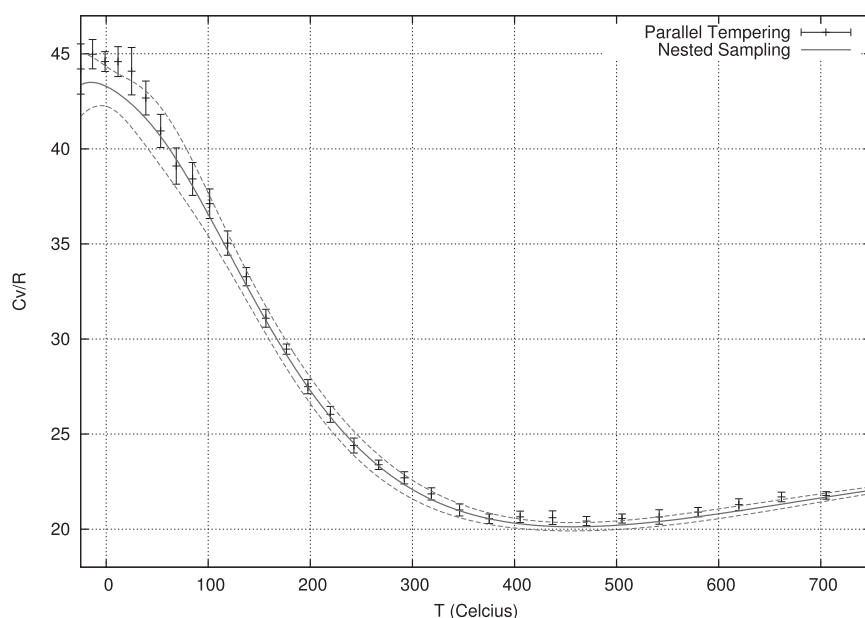


FIGURE 2 Heat capacity curve for the 16-residue polyalanine. The nested sampling simulations (solid line) use 1.3×10^9 MC moves, with error lines denoting two standard errors from the mean. The parallel tempering uses the same number of MC moves again with error bars showing two standard errors from the mean.

(ubiquitin-like) fold, which has been extensively studied by a variety of folding simulation techniques (37–40). Its native structure is shown on the left of Fig. S3. All figures of protein G in this article have been oriented so that the first β -strand is the second strand from the right and the N-terminal residue is at the top.

As described above, the nested sampling procedure can be used to estimate the thermodynamic energy of the system at any temperature, using Eq. 4. For protein G, at room temperature ($\beta = 1.0$), the thermodynamic energy is -190 units. Fig. S3 shows a sample of four room-temperature, thermodynamically accessible conformations found by a single nested sampling simulation with $K = 20,000$ and $m = 15,000$. The conformers have energies -189 , -190 , -191 , and -190 , respectively, with backbone RMSDs (from the crystal structure) of 1.93 Å, 2.96 Å, 3.97 Å, and 5.22 Å, respectively. The estimated value of the backbone RMSD at $\beta = 1$, calculated using Eq. 5, is $\mathbb{E}(\text{RMSD}|\beta = 1) = 3.21$ Å.

Conformers A–D in Fig. S3 have the correct backbone topology, close to the native structure, but there is a reasonable amount of variation in the orientation of the helix with respect to the β -sheet at this temperature. It is important to remember that protein structures are intrinsically flexible (41–43), and the crystal structure (1PGA.pdb) is only one member of an ensemble of conformations that the protein may explore. In the Supporting Material we demonstrate that flexible motion of protein G allows a substantial re-orientation of the axis of the helix with respect to the sheet. Conformers A–D in Fig. S3, which differ from the native state principally in the orientation of the helix relative to the sheet, may therefore be more representative of the native state than the RMSD alone suggests.

The first half of the nested sampling simulation is spent exploring high-energy conformations with no noticeable secondary structure and often steric hindrances. In the second half of the simulation, once the long-range quadratic bias potential has pulled the secondary structure elements close together, the short-range hydrogen-bond interaction contributions increase to dominate the bias potential contributions, having a steeper gradient in the last third of the simulation (see Fig. S4 (top)). The short-range hydrophobic interaction contributions are the smallest, but nevertheless not negligible; they ensure the correct packing of the hydrophobic and amphipathic side chains at conformations available at room temperature (see below and Fig. 3). Fig. S4 (bottom) shows a sequence of 10 conformers in order of decreasing energy. These conformers come from the deepest basin of the energy landscape chart of a simulation (higher energy conformers come from the part of the energy landscape chart that contains the deepest basin). The sequence illustrates how the secondary and tertiary structure accrete in the course of a simulation, capturing the essence of the hierarchical folding model. The sequence is not, however, a single folding pathway, in the sense of a molecular dynamics trajectory; there are many conformations in the

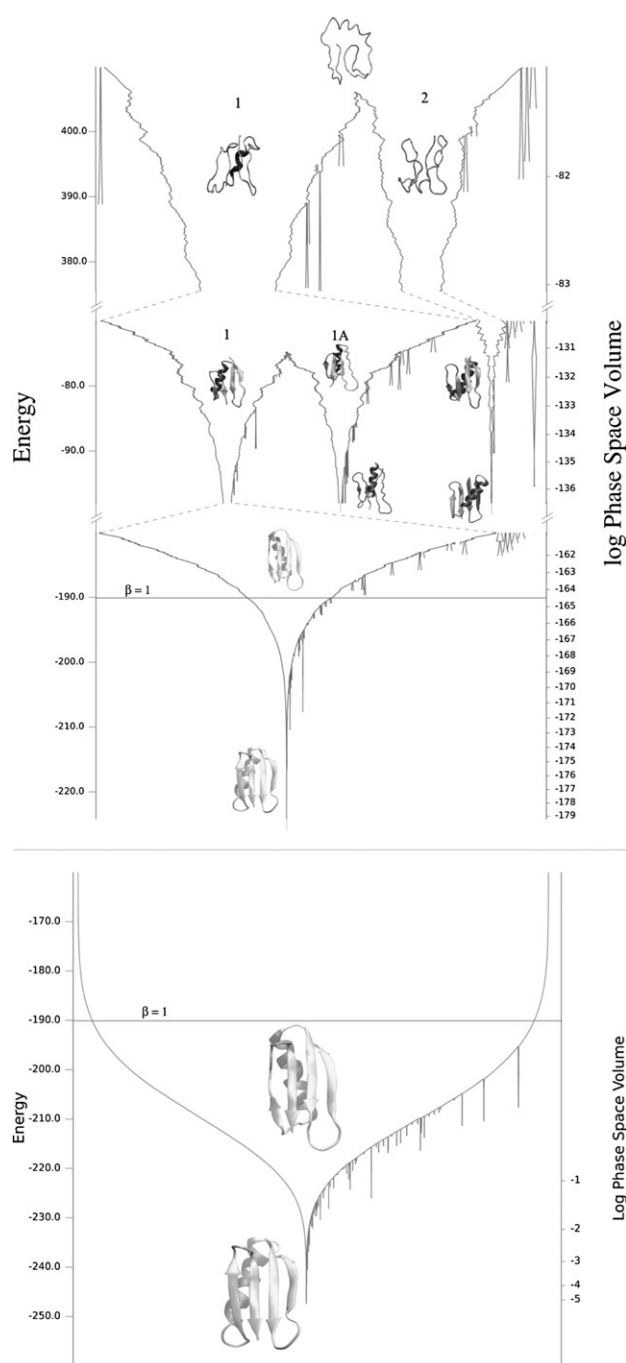


FIGURE 3 (Top) Prior (potential) energy landscape chart. (Bottom) Posterior energy landscape chart at $\beta = 1$, for a nested sampling simulation of protein G using $K = 20,000$ and $m = 15,000$. (Left axis) The energy is shown in units of RT , and the width of the chart is proportional to the sum of the prior (top) and posterior (bottom) weights of the nested sampling points below the given energy level (shown on the right axis). The prior energy landscape chart shows the potential energy surface and the posterior energy landscape chart shows, for a given temperature, the probabilities of finding conformations from the different basins. At $\beta = 1$ (room temperature), only funnel 1 is accessible. The scaling function used for the prior energy landscape chart is $\exp(-fE)$ (with f being 0.1 , 0.4 , and 0.7 on the top, middle, and bottom panels of the prior energy landscape chart, respectively). Example conformers from the main basins, at various energy levels, are shown on the charts.

active set in the same basin. It is possible, though, that Fig. S4 (bottom) represents a plausible sequence of events leading to the native structure.

Energy landscape charts using the prior and posterior weights for a nested sampling simulation of protein G using $K = 20,000$ and $m = 15,000$, calculated using a connectivity number $k = 15$, are shown in Fig. 3. The volume scale on the right-hand axis shows the proportion of the prior and posterior phase space volume available below the given energy level. The width of the chart uses this scale. Basins that contain $<1/1000$ th of the probability mass at the point of splitting are not shown on the diagram. Conformers have been placed on the chart to provide examples of the samples found in different places of the chart.

Topologically, for energy above 405 units, there is one main basin containing virtually all of the samples. There is little structure in the samples, as shown by the conformer at the top of the chart. However, at energy 405 units, the phase space splits into two main funnels: one with the helix forming on the correct side of the sheet (funnel 1) and one with it forming on the incorrect side (funnel 2). Funnel 1 further splits at energy -75 units, corresponding to conformations where the hydrophobic residues are in the interior of the protein (funnel 1) or on the surface of the protein (funnel 1A). At room temperature (the expected energy corresponding to $\beta = 1$ is marked by a horizontal line on both panels of the chart), the phase space volume of both funnels 1A and 2 are $<1/1000$ th of the main funnel and hence the posterior energy landscape chart consists of a single funnel. The inaccessibility of funnel 1A at room temperature indicates the importance of hydrophobic interactions.

Fig. S5 shows two conformers that are placed in the same small basin, branching off the right-hand funnel. The conformer on the left has higher energy than the one on the right. These conformers are very similar, and demonstrate that the topological analysis shows how metastable conformations are formed. The pathway to these states would be obtained by considering conformers found in the same basin.

DISCUSSION

It is interesting to consider how the energy landscape charts vary from simulation to simulation. Topologically, we always find two main funnels in the protein G simulations (funnels 1 and 2 on Fig. 3), corresponding to the packing of the helix on either side of the sheet. The dominant mode with the native like backbone topology (funnel 1) splits again at a lower energy level to two funnels, corresponding to the hydrophobic residues being in the interior (funnel 1) or on the surface (funnel 1A) of the protein. The energy at which funnels 1 and 2 split varies significantly between simulations, from 220 to 580 energy units. This is probably because the RMSD metric is an imprecise way of comparing wildly different conformations. The energy

where funnels 1 and 1A split has a much smaller variation, -75 to -55 energy units. This trend in the variation of splitting energies was also observed in the nested sampling simulations of the other modeled proteins. Metrics other than the RMSD might improve the reproducibility of energy landscape charts and would be worthy of investigation.

The relative basin widths of energy landscape charts depend on the size of the nested sampling active set, K . In general, K determines the resolution of exploration. When converging the evidence at lower temperatures, a larger value of K is required. This is because at every splitting of the likelihood function, the probability of exploring the dominant mode decreases, according to Eq. 3. At high energies, the accessible conformational space is connected, and the MMC procedure explores the space effectively. As the energy lowers, the accessible conformational space becomes increasingly disconnected. Because the MMC procedure cannot jump between disconnected components of the conformational space, an increasingly large set of active points is required to sample effectively. As the posterior mass is concentrated at lower energies for lower temperatures, K behaves as an effective minimum temperature. Using too small an active set for a given temperature causes large variation between different nested sampling simulations; for example, the estimates for the evidence and the relative widths of the funnels of energy landscape charts.

In the protein G simulations, we find that $K = 20,000$ is large enough to produce simulation independent charts for temperatures near $\beta = 1$. When using, for example, $K = 2500$, which is too small for sampling the posterior distribution at $\beta = 1$, we find that the active set becomes extremely homogenous and the simulation is, in effect, exploring just one tiny basin in one of the main funnels, by making smaller and smaller crankshaft rotations. Hence, we find a single room-temperature accessible conformation, as opposed to the wide selection that is found when $K = 20,000$.

The magnitude of m relative to K is problem-specific. It has been suggested that for probability distributions that lack a large number of modes, it is optimal to set K small and use a large m (the cost is proportional to mK) (21). For protein G, we find the energy landscape is so complex that we need a large K to explore all the funnels simultaneously, and a large m to ensure the active set remains heterogeneous, and we therefore choose m and K to have the same order of magnitude. Incorporating nonlocal flexible motions (44) into our MMC procedure may allow a decrease in m without losing heterogeneity and this is a focus of future work. If this proves to be the case, we would choose to increase K relative to m .

In our previous work (27), using MMC with parallel tempering to simulate the folding of protein G with a simpler model (no γ -atoms and hydrophobic interactions were included in this model), the lowest energy structures obtained were similar to those shown at the bottom of funnel 2 of Fig. 3, with the helix packed on the incorrect side of

the sheet and a backbone RMSD of 8.6 Å from the crystal structure. This demonstrates the difficulty of using parallel tempering or simulated annealing to reconstruct the native structure, when the energy landscape exhibits two main funnels separated by a large energy barrier. If the annealing proceeds down the incorrect funnel it will be nearly impossible for it to climb back out and down into the correct funnel.

The reason for the double funnel is the symmetry of the protein G topology with respect to the Gō-type bias potential, which is the predominant factor at the beginning of the simulation. The further splitting of the main funnel into funnels 1 and 1A (Fig. 3) is also due to the nature of the Gō-type bias potential. This applies a quadratic potential on the C_β atom contacts, which does not restrict the hydrogen-bond pattern between the individual strands; at high energies, both conformations (with the hydrophobic residues of the β -sheet being in the interior or on the surface of the protein) are similarly likely to be adopted. However, other energy and entropy contributions due to the presence of side chains (e.g., hydrophobic interactions and steric clashes) ensure that only conformations with the natively like topology are accessible at room temperature. This way, energy landscape charts also reflect the nature of the protein model and force field used. For example, the energy landscape charts for chymotrypsin inhibitor 2, which differs in topology from protein G, but also possesses a similar symmetry with regard to the packing of the α -helix against the β -sheet, also exhibit this double funnel (see the [Supporting Material](#)).

It would be interesting to compare energy landscape charts of nested sampling simulations using other protein models and force fields, for example, all-atom representations, and this will be a focus of future work.

CONCLUSION

This article has described the parallelization of the nested sampling algorithm, and its application to the problem of protein folding in a force field of empirical potentials that were designed to stabilize secondary structure elements in room-temperature simulations. The output of the nested sampling algorithm can be used to produce energy landscape charts, which give a high level description of the potential energy surface for the protein folding simulations. These charts provide qualitative insights into both the folding process and the nature of the model and force field used. The topology of the protein molecule emerges as a major determinant of the shape of the energy landscape, as has been noted by other authors (37). The energy landscape chart for protein G exhibits a double funnel with a large energy barrier, a potential energy surface that parallel tempering struggles to explore fully. The nested sampling algorithm also provides an efficient way to calculate free energies and the expectation value of thermodynamic

observables at any temperature, through a simple postprocessing of the output.

SUPPORTING MATERIAL

Supporting text with 11 figures and references (45–62) is available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(12\)00055-0](http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)00055-0).

We thank Gábor Csányi, Konrad Dabrowski, Livia Pártay, Alexei Podtelnikov, and John Skilling for helpful discussions.

We acknowledge support from the Leverhulme Trust (grant F/00 215/BL to N.S.B., C.V., and D.L.W.) and the Engineering and Physical Sciences Research Council, UK (grant EP/G021163/1 to D.L.W.). S.A.W. acknowledges support from the Leverhulme Trust (Early Career Fellowship).

REFERENCES

1. Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science*. 181:223–230.
2. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*. 84:7524–7528.
3. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
4. Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24:26–33.
5. Baldwin, R. L., and G. D. Rose. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* 24:77–83.
6. Lindorff-Larsen, K., S. Piana, ..., D. E. Shaw. 2011. How fast-folding proteins fold. *Science*. 334:517–520.
7. Gō, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.
8. Takada, S. 1999. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. USA*. 96:11698–11700.
9. Dinner, A. R., A. Sali, ..., M. Karplus. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25:331–339.
10. Sali, A., E. Shakhnovich, and M. Karplus. 1994. How does a protein fold? *Nature*. 369:248–251.
11. Clementi, C., A. E. García, and J. N. Onuchic. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.* 326:933–954.
12. Hori, N., G. Chikenji, ..., S. Takada. 2009. Folding energy landscape and network dynamics of small globular proteins. *Proc. Natl. Acad. Sci. USA*. 106:73–78.
13. Amato, N. M., K. A. Dill, and G. Song. 2003. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.* 10:239–255.
14. Becker, O. M., and M. Karplus. 1997. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* 106:1495–1517.
15. Wales, D. J. 2004. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*. Cambridge University Press, Cambridge, UK.
16. Wales, D. J., and T. V. Bogdan. 2006. Potential energy and free energy landscapes. *J. Phys. Chem. B*. 110:20765–20776.
17. Skilling, J. 2006. Nested sampling for general Bayesian computation. *J. Bayesian Anal.* 1:833–860.
18. Sivia, D. S., and J. Skilling. 2006. *Data Analysis, A Bayesian Tutorial*, 2nd ed. Oxford University Press, Cambridge, UK.

19. Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*. 220:671–680.
20. Wang, F., and D. P. Landau. 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86:2050–2053.
21. Swendsen, R. H., and J. S. Wang. 1986. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* 57:2607–2609.
22. Murray, I., D. MacKay, ..., J. Skilling. 2006. Nested sampling for Potts models. *Adv. Neural Inf. Process. Syst.* 18:947–954.
23. Pártay, L. B., A. P. Bartók, and G. Csányi. 2010. Efficient sampling of atomic configurational spaces. *J. Phys. Chem. B*. 114:10502–10512.
24. Bennett, C. H. 1976. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* 22:245–268.
25. Mukherjee, P., D. Parkinson, and A. R. Liddle. 2006. A nested sampling algorithm for cosmological model selection. *J. Astrophys. Lett.* 638:51–54.
26. MacKay, D. J. C. 2004. www.inference.phy.cam.ac.uk/bayesys/box/nested.pdf.
27. Podtelezhnikov, A. A., and D. L. Wild. 2009. Reconstruction and stability of secondary structure elements in the context of protein structure prediction. *Biophys. J.* 96:4399–4408.
28. Podtelezhnikov, A. A., and D. L. Wild. 2008. CRANKITE: a fast polypeptide backbone conformation sampler. *Source Code Biol. Med.* 3:12.
29. Podtelezhnikov, A. A., Z. Ghahramani, and D. L. Wild. 2007. Learning about protein hydrogen bonding by minimizing contrastive divergence. *Proteins*. 66:588–599.
30. Podtelezhnikov, A. A., and D. L. Wild. 2005. Exhaustive Metropolis Monte Carlo sampling and analysis of polyaniline conformations adopted under the influence of hydrogen bonds. *Proteins*. 61:94–104.
31. Srinivasan, R., and G. D. Rose. 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci. USA*. 96:14258–14263.
32. Creamer, T. P., and G. D. Rose. 1995. Interactions between hydrophobic side chains within α -helices. *Protein Sci.* 4:1305–1314.
33. Luom, P., and R. L. Baldwin. 2002. Origin of the different strengths of the (I_{i+4}) and (I_{i+3}) leucine pair interactions in helices. *J. Biophys. Chem.* 96:103–108.
34. Elofsson, A., S. M. Le Grand, and D. Eisenberg. 1995. Local moves: an efficient algorithm for simulation of protein folding. *Proteins*. 23:73–82.
35. Sibanda, B. L., T. L. Blundell, and J. M. Thornton. 1989. Conformation of β -hairpins in protein structures. A systematic classification with applications to modeling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* 206:759–777.
36. Ferrenberg, A. M., and R. H. Swendsen. 1989. Optimized Monte Carlo analysis. *Phys. Rev. Lett.* 63:1195–1198.
37. Karanicolas, J., and C. L. Brooks, 3rd. 2002. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* 11:2351–2361.
38. Kolinski, A., and J. Skolnick. 2004. Reduced models of proteins and their applications. *Polymer (Guildf.)*. 45:511–524.
39. Sheinerman, F. B., and C. L. Brooks, 3rd. 1998. Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.* 278:439–456.
40. Shimada, J., and E. I. Shakhnovich. 2002. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci. USA*. 99:11175–11180.
41. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. *Nature*. 450:964–972.
42. Thorpe, M., M. Lei, ..., L. A. Kuhn. 2001. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.* 19:60–69.
43. Wells, S., S. Menor, ..., M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* 2:S127–S136.
44. Jimenez-Roldan, J. E., R. B. Freedman, ..., S. A. Wells. 2012. Protein flexibility explored with normal modes and geometric simulation. *Phys. Biol.* In press.
45. Engh, R. A., and R. Huber. 1991. Accurate bond and angle parameters for x-ray protein structure refinement. *Acta Crystallogr. A*. 47:392–400.
46. Engh, R. A., and R. Huber. 2001. Structure quality and target parameters. In *International Tables for Crystallography*, Vol. F: Crystallography of Biological Macromolecules, 1st ed. M. G. Rossman and E. Arnold, editors. Kluwer Academic Publishers for the International Union of Crystallography, Dordrecht, Boston, London. 382–392.
47. Brünger, A. 1992. X-PLOR, V. 3.1: A System for X-Ray Crystallography and NMR. Yale University Press, New Haven, CT.
48. Ho, B. K., E. A. Coutasias, ..., K. A. Dill. 2005. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* 14:1011–1018.
49. Brenner, S. E., P. Koehl, and M. Levitt. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* 28:254–256.
50. Carreira-Perpinan, M., and G. Hinton. 2005. On contrastive divergence learning. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, Jan. 6–8, 2005, The Savannah Hotel, Barbados. 217.
51. Ramachandran, G. N., C. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7:95–99.
52. Hopfinger, A. J. 1973. *Conformational Properties of Macromolecules*. Academic Press, New York.
53. Word, J. M., S. C. Lovell, ..., D. C. Richardson. 1999. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* 285:1711–1733.
54. Pappu, R. V., R. Srinivasan, and G. D. Rose. 2000. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl. Acad. Sci. USA*. 97:12565–12570.
55. Berman, H. M., J. Westbrook, ..., P. E. Bourne. 2000. The Protein DataBank. *Nucleic Acids Res.* 28:235–242.
56. Baker, E. N., and R. E. Hubbard. 1984. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* 44:97–179.
57. Savage, H. J., C. J. Elliott, ..., J. M. Finney. 1993. Lost hydrogen bonds and buried surface area: rationalizing stability in globular proteins. *J. Chem. Soc., Faraday Trans.* 89:2609–2617.
58. Stickle, D. F., L. G. Presta, ..., G. D. Rose. 1992. Hydrogen bonding in globular proteins. *J. Mol. Biol.* 226:1143–1159.
59. McDonald, I. K., and J. M. Thornton. 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238:777–793.
60. Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.
61. Word, J. M., S. C. Lovell, ..., D. C. Richardson. 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* 285:1735–1747.
62. Suhre, K., and Y.-H. Sanejouand. 2004. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.* 32(Web Server issue):W610–W614.